Rates and timing of chlorophyll-α increases and related environmental variables in global temperate and cold-temperate lakes

Hannah Adams¹, Jane Ye¹, Bhaleka Persaud¹, Stephanie Slowinski¹, Homa Kheyrollah Pour², Philippe Van Cappellen¹

¹Ecohydrology Research Group, Department of Earth and Environmental Sciences and Water Institute, University of Waterloo, Waterloo, ON, Canada
²ReSEC Research Group, Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, ON, Canada

Correspondence to: Hannah Adams (hadams21@mun.ca)

Running Head: Chlorophyll-α concentrations in northern lakes

Keywords: northern lakes, primary productivity, chlorophyll-α concentration, increase rate, trophic state, climate change
Abstract
Lakes are key ecosystems within the global biogeosphere. However, the environmental controls on the biological productivity of lakes, including surface temperature, ice phenology, nutrient loads and mixing regime, are increasingly altered by climate warming and land-use changes. To better characterize global trends in lake productivity, we assembled a dataset on chlorophyll-a concentrations, as well as associated water quality parameters and surface solar radiation, for temperate and cold-temperate lakes experiencing seasonal ice cover. We developed a method to identify periods of rapid net increase of in situ chlorophyll-a concentrations from time series data and applied it to data collected between 1964 and 2019 across 343 lakes located north of 40°. The data show that the spring chlorophyll-a increase periods have been occurring earlier in the year, potentially extending the growing season and increasing the annual productivity of northern lakes. The dataset on chlorophyll-a increase rates and timing can be used to analyze trends and patterns in lake productivity across the northern hemisphere or at smaller, regional scales. We illustrate some trends extracted from the dataset and encourage other researchers to use the open dataset for their own research questions.
1 Introduction

Lakes play an important role in the biogeochemical cycling of many elements (Battin et al., 2008; Cole et al., 2007; O’Connell et al., 2020; Rousseaux and Gregg, 2013; Schindler, 1971). With over 100 million documented lakes on earth (Verpoorter et al., 2014), evidence indicates that the majority of global lakes are shallow with enough light and nutrients available to make them highly productive ecosystems (Downing et al., 2006; Wetzel, 2001). Lakes therefore represent active sites for the storage, transport, and transformation of carbon, nutrients (e.g., nitrogen, phosphorus, silicon, iron), and contaminants (e.g., mercury) along the freshwater continuum (Lauerwald et al., 2019; Tranvik et al., 2009). They are also sensitive to the effects of climate change (Williamson et al., 2009; Rouse et al., 1997).

There are multiple environmental controls on lake primary productivity, including water temperature, ice phenology, nutrient concentrations, circulation, mixing regime, and solar radiation (Lewis, 2011; Zohary et al., 2009). Stressors such as climate change and nutrient pollution can significantly impact these controls, altering the ecosystem structure and biogeochemical functioning of lakes (Jeppesen et al., 2020; Markelov et al., 2019). Changes affecting northern lakes include warmer water temperatures, enhanced stratification and hypoxia, nutrient enrichment, light attenuation by chromophoric organic matter, and increases in the relative abundance of toxic cyanobacteria in the phytoplankton community (Deng et al., 2018; Huisman and Hulot, 2005; Jeppesen et al., 2003; Creed et al., 2018). For example, Lake Superior has seen an increase in primary production during the last century, together with increasing surface water temperatures and longer seasonal stratification and ice-free periods (O’Beirne et al., 2017). Other lakes are similarly experiencing increases in productivity. According to Lewis (2011), the current mean primary production of lakes is 260 g C m⁻² y⁻¹, which is 162% higher than earlier estimations under historical baseline conditions.

Globally, phytoplankton (i.e., algae) are the main primary producers in lakes and generally make up the foundation of lentic food webs (Carpenter et al., 2016). Periods of high lake productivity coincide with a rapid increase in phytoplankton biomass. In extreme cases, algal blooms can reach hundreds to thousands of cells per milliliter (Henderson-Seller and Markland, 1987). These bloom events produce large quantities of decomposing organic matter that cause the expansion of hypoxic conditions within the lake (Watson et al., 2016). In harmful algal blooms, certain
algal species also release hepatotoxic and neurotoxic compounds (Codd et al., 2005). Thus, identifying trends in the timing and intensity of seasonal algal growth, and linking them to changes in environmental stressors, can help predict the future of lake productivity and assess the risk of undesirable algal blooms.

Because it is challenging to measure algal abundance and growth directly, chlorophyll-α is often used as a proxy for algal biomass and an indicator of the associated primary production in lakes (Huot et al., 2007). Although other proxies have been developed (Lyngsgaard et al., 2017), chlorophyll-α is the most common metric to characterize trends in algal biomass within and across lakes, especially in historical water quality records. Tett (1987) proposes a chlorophyll-α threshold of 100 µg L⁻¹ to define “exceptional blooms”, Jonsson et al. (2009) use a threshold of 5 µg L⁻¹ to identify a bloom, while Binding et al. (2021) flags an algal bloom when the chlorophyll-α concentrations extracted from satellite observations exceed 10 µg L⁻¹. Such threshold values, however, do not take into account the baseline (i.e., no-bloom) chlorophyll-α concentration specific to a given lake, or the lake’s trophic status (Germán et al., 2017). Furthermore, focusing on harmful and nuisance algal blooms alone may mask the impact that a changing climate or other stressors may have on a lake’s overall biological productivity.

Intra-annual fluctuations in lake chlorophyll-α concentration result from the interactions of multiple variables and processes including grazing by zooplankton, competition between algal species with different growth strategies and chlorophyll-α contents, and changes in temperature, light, and nutrient availability (Lyngsgaard et al., 2017; Sommer et al., 1986). In dimictic lakes, for example, there are usually two peaks in algal biomass, and hence also in chlorophyll-α concentrations, in the spring and fall, with a smaller biomass stock of slower growing species during the summer, and an even smaller stock of algae (in terms of both biovolume and chlorophyll-α) under the ice cover in the winter (Hampton et al., 2017).

The spring increase in algal biomass generally consists of fast-growing algal species that take advantage of the increases in temperature and light following ice-off, as well as the available inorganic nutrients that were generated by mineralization under the ice over the winter. The shift from spring to summer algal communities often coincides with high zooplankton grazing rates exceeding the spring algal growth rates, hence, bringing down the total algal biomass. The high zooplankton grazing rates favor the growth during the summer of algal species that are less
edible by grazers, but which tend to grow at slower rates. Lake overturn in the fall initiates the
transition from the predominance of the slow growing species in the summer to the fast-growing
phytoplankton species in the fall causing a second peak in algal biomass (Sommer et al., 1986).

A common approach for comparing chlorophyll-a trends across multiple lakes is to consider the
maximum or mean annual chlorophyll-a concentrations. For example, Ho et al. (2019)
100 applied the Mann-Kendall trend test to analyze time series of annual maximum
chlorophyll-a concentrations, while Shuvo et al. (2021) used a random forest
regression approach to assess the relative importance of climatic versus non-climatic controls on
mean chlorophyll-a concentrations. Both these studies analyzed chlorophyll concentrations
derived from satellite observations rather than measured in situ. In addition, these approaches did
not specifically identify the periods of the year when chlorophyll-a concentrations experienced
rapid changes.

Alternatively, the rate of increase in chlorophyll-a concentration can be used to constrain the
timing of rapid increase in algal biomass usually associated with periods of high primary
productivity. In this study, we refer to these as “periods of chlorophyll-a increase” (PCIs). The
weeks leading up to a PCI are crucial to create the necessary conditions that enable algal growth
(Lewis et al., 2018). Thus, to analyze trends in lake net primary productivity, one should
consider environmental variables, such as surface water temperature, solar radiation, and nutrient
concentrations, both during and preceding the annual PCIs.

Although the rate of chlorophyll-a concentration increase has been used to detect algal blooms
within individual water bodies, for example in the San Roque reservoir (Germán et al.,
2017), it has rarely been used across large temporal (i.e., more than a few
years) and spatial (i.e., regional and up) scales. Here, we present a method for calculating net
rates of chlorophyll-a increase (RCI). The timing of PCIs and values of the corresponding RCIs
were derived from in situ chlorophyll-a concentrations obtained for 343 lakes located at latitudes
above 40° N. The entire dataset covers the period from 1964 to 2019, and further contains data
on coincident environmental control variables, including surface solar radiation. To illustrate the
potential applications of the resulting dataset, we present some temporal trends of the
chlorophyll-a rates and their relationships with environmental variables. The dataset is made
available as an open resource that other researchers are encouraged to use in their own work.
2 Data and methods

All data processing, visualizations, and analyses were carried out with Python (ver. 3.7.6; Python Software Foundation, 2021) using the pandas library (Reback et al., 2020), NumPy library (Harris et al., 2020), and Dplython library (Riederer, 2015), while QGIS/PYQGIS was used for all spatial data analyses (ver. 3.16; QGIS Development Team, 2021).

2.1 Data acquisition, compilation, and quality control

2.1.1 Lake data selection

*In situ* chlorophyll-α concentrations and other lake physico-chemical data were extracted from open source international, national, and regional databases (see Table A1 for a summary of all databases used). The data include surface water temperature, Secchi depth and pH, as well as the concentrations of particulate organic carbon (POC), total phosphorus (TP), soluble reactive phosphorus (SRP), total Kjeldahl nitrogen (TKN) and dissolved organic carbon (DOC).

To enable readers to compare the methods used by different lake monitoring agencies and researchers to collect and process *in situ* samples, we provide the links to the raw data sources and metadata files in the appendix (Tables A1-A3). When selecting data, we were as consistent as possible by implementing the following steps (more details can be found in the “initial formatting” folder found in the associated GitHub repository).

1) We only included measurements taken at ≤ 3 m water depth. When the sampling depth was not provided, we assumed the sample was taken from within the top 0.5-3 m of the lake, given that this is the usual standard sampling protocol (Dorset Environmental Science Centre, 2010; United States Environmental Protection Agency, 2012).

2) We selected lakes from mid-to-high latitudes (≥ 40° N). Lakes at these latitudes typically experience seasonal ice cover and thermal stratification during the summer, in contrast to low-latitude lakes that are typically meromictic or polymictic (Woolway and Merchant, 2019).

We omitted all variable values below the corresponding analytical detection limit. Data from different sources were individually reformatted to yield consistent (standard) units and headings. Where needed, reported values were averaged to yield daily mean values before being combined.
into a single csv file. When multiple chlorophyll-\( \alpha \) data types were available (as, for example, in the Laurentian Great Lakes data series), we selected the uncorrected data because most reported lake chlorophyll-\( \alpha \) concentrations have not been corrected for phaeophytin pigments. If no coordinates were provided, we assigned those of the lake centroid in QGIS. Fifteen lakes had unknown location and were removed from the final dataset. We further restricted ourselves to lakes that in most years were sampled at least 6 times per year, which was considered the minimum sampling frequency to reliably detect the yearly PCIs. Lake names were standardized by expanding on abbreviations and removing unnecessary capitalization and special characters.

With the above selection criteria, the final dataset contained 52116 potential PCIs for 343 lakes at \( \geq 40^\circ\text{N} \) and covering the period 1964-2019. The location of the lake sampling locations in the PCI dataset are shown in Figure 1.

2.1.2 Surface solar radiation data

Open source \textit{in situ} surface solar radiation (SSR) data for the period 1950-2020 were collected from stations paired with the selected lakes (see Table A2 for data sources). Each lake was paired with the closest SSR station using the nearest neighbor function in QGIS, allowing for a maximum radius of three degrees (Schwarz et al., 2018; Figure 1). In the dataset provided here, the geodesic distance between each lake and its paired SSR station is given, as well as the difference in elevation.

The SSR data temporal resolution varied from minutes to months. Hence, where needed, the SSR data were resampled to yield monthly mean values. For the Experimental Lakes Area (ELA) in Ontario, Canada, the data were converted from photosynthetically active radiation (PAR) to SSR, where the PAR wavelength range (400-700 nm) was averaged to 550 nm.

2.1.3 Lake characteristics

For each lake, we calculated the trophic status index (TSI) based on the mean chlorophyll-\( \alpha \) concentration over the sampling period. This TSI value was used to assign the lake to the corresponding trophic state category according to Carlson and Simpson (1996). The HydroLAKES shapefile yielded the lake’s surface area, mean depth, and volume (ver. 1.0; Messager et al., 2016). Lake elevation was extracted from a digital elevation model (DEM) (Danielson and Gesch, 2010), and each lake was assigned its corresponding climate zone using HydroATLAS data (ver. 1.0; Linke et al., 2019). The metadata for these variables are published...
as part of the data publication (Adams et al., 2021), and a summary table of associated lake data is provided in the appendix (Table A4).

2.2 Detecting seasonal periods of chlorophyll-α increase

Periods of chlorophyll-α increase (PCIs) were identified based on the normalized net rate of change in chlorophyll-α concentration (NRCC) at each lake sampling point throughout the year. To locate the start and end of a PCI, we smoothed the annual chlorophyll-α time series using a Savitzky-Golay filter (SciPy.signal.savgol_filter) and flagged optima in the smoothed data (SciPy.signal.find_peaks) using functions from the open source SciPy ecosystem (Virtanen et al., 2020). The procedure is illustrated in Figure 2.

The NRCC at any given time during the year was calculated by computing the first derivative of the smoothed chlorophyll-α concentration versus time and dividing the derivative value by the corresponding chlorophyll-α concentration. For each lake and each year, the start of the first PCI was defined as the day the NRCC surpassed 0.4 day⁻¹. This threshold rate was selected following a series of sensitivity tests (details provided in the supplementary information). A threshold NRCC value was considered preferable than a threshold RCI value because it accounts for variations among lakes and among years in the baseline chlorophyll-α concentrations during the non-growing season.

The PCI ended on the day the peak in chlorophyll-α concentration was reached, that is, just before the NRCC turned negative. If a threshold NRCC of 0.4 day⁻¹ was not reached during a given year, the PCI began when the NRCC first became positive. The second (fall) PCI was identified in the same way, following the end of the first (spring) PCI. If the annual chlorophyll-α concentration only yielded one peak value in the smoothed data series, only one PCI was identified for that year, which was then labelled as a “single PCI” year. Years with more than two chlorophyll-α peaks or with no peaks were not included in the PCI dataset.

Depending on data availability, the pre-PCI period was defined as the one- or two-week period immediately preceding the PCI start day. For each pre-PCI, the mean surface water temperature, SSR, and TP concentration were compiled. These served as simple indicators of how favorable in-lake conditions were to initiate algal growth (Lyngsgaard et al., 2017). An example of a year
with a spring and fall PCI is shown in Figure 3. Note that we use the label “fall” to indicate the second yearly PCI, although in some cases the fall PCI was initiated before the fall equinox.

Once the PCI and pre-PCI durations were determined, the mean values of the variables listed in Table 1 were calculated. This was done for each lake and for each year data were available. In the dataset, each row represents a single PCI and includes the timing and duration, RCI value, plus the mean values for all other relevant lake variables, including SSR, averaged for the PCI and pre-PCI. Note that, along with the variables in Table 1, we included the total number of samples collected each year and the mean time between samples. Thus, if desired the user can filter the dataset for a higher sampling frequency than done here. The supplementary information of the dataset also identifies the organization responsible for monitoring each lake.

3 Dataset: data distributions

3.1 Dataset characteristics

Most lakes in the dataset are located between 50 and 60° N. The majority of available open data are from organizations within the United Kingdom, Sweden, Canada, and the United States. The years with available data in the dataset are unevenly distributed. The majority of PCIs fall in the period 2005-2019 (Figure 4a), likely due to a combination of increased lake monitoring efforts and a push in recent years towards greater accessibility of publicly funded data (Hallegraeff et al., 2021; Roche et al., 2020). Most sampling frequencies are in the range of 25 to 30 days, with additional peaks at 7 and 14 days (Figure 4b). Thus, with a few exceptions, the PCIs included in the dataset occurred in lakes sampled at a monthly frequency or better.

The distribution of trophic states of the PCIs recorded in the dataset are: 1.6% oligotrophic, 18.6% mesotrophic, 75.2% eutrophic, and 4.6% hypereutrophic. Single PCIs dominate oligotrophic lakes where they make up 96.1% of all PCIs (Figure 4c). This may reflect the severe nutrient limitation in oligotrophic lakes, which prevents the occurrence of a second annual algal PCI (Rigosi et al., 2014). Oligotrophic lakes also tend to dominate at latitudes ≥ 55 °N (Figure 4d) where lower water temperatures and lower cumulative solar radiation may further limit algal growth (Lewis, 2011). The PCI durations range from 3 to 275 days, with a median of 68 days (Figure 5a). Fall PCIs tend to be shorter than spring and single PCIs, with the latter exhibiting the most variable start and end days (Figure 5b).
3.2 Environmental conditions during PCIs

Rates of chlorophyll-a increase during the PCIs exhibit log-normal distributions (Figure 6a). The mean chlorophyll-a rate is lowest in the single PCI category and highest in the fall PCIs. Mean surface water temperature has a distinct bimodal spring-fall distribution (Figure 6b). For the single PCIs, the corresponding mean temperatures are evenly distributed across the annual range, which reflects the large spread in the timing of the single PCIs (Figure 5b). Total P concentrations are lowest during the spring PCIs (Figure 6c), consistent with a greater control of P limitation on algal growth during spring compared to summer and fall (Kirillin et al., 2012). Secchi depth during the PCIs ranges from 0.01 to 15.4 m, with fall PCIs experiencing the lowest mean Secchi depth (Figure 6d), as turbidity generally increases after the spring bloom.

4 Dataset: examples of trends

The PCI delineation and the estimation of RCI can in principle be applied to any lake for which time series chlorophyll-a concentration data are available. By creating a dataset comprising many lakes and covering multi-year time periods, it becomes possible to extract global trends in lake chlorophyll-a. Here, we provide a few illustrative examples of how the dataset can be interrogated, setting the stage for its use and extension by other researchers.

4.1 Chlorophyll-a rates: trophic status, latitude and climate zone

When grouped by trophic status, mean and median chlorophyll-a growth rates (RCIs) show the expected increase from oligotrophic to hypereutrophic lakes (Figure 7a). The rates in the different trophic categories, however, cover large and overlapping ranges. When grouped according to latitude, lakes between 40 and 50° N exhibit the widest range in RCIs (Figure 7b), in part due to the high proportion of lakes in this latitude range. The highest latitude lakes (60-70° N) tend to have the lowest RCIs, which may reflect the cooler temperatures experienced (Lewis, 2011).

The lakes are spread across three climate zones: cold and mesic; cool, temperate, and dry; and warm, temperate, and mesic (Figure 7c). There is considerable overlap in RCI across the climate zones, with no systematic differences in the mean and median RCI values between the zones.

While variations in chlorophyll-a rates of increase (RCIs) are often assumed to reflect comparable differences in algal biomass growth rates, it is important to note that the chlorophyll-
a to biomass ratio varies within and among lakes. In particular, chlorophyll-\(a\) to biomass ratios are known to be sensitive to variations in solar radiation, temperature, algal species, and cell size (Baumert and Petzoldt, 2008; Inomura et al., 2019; Geider, 1987; Álvarez et al., 2017). The summer ratio of chlorophyll-\(a\) to biomass (the latter typically expressed as particulate organic carbon concentration) generally increases with increasing latitude because algae are adapted to harvest the more variable daylight conditions, including longer summer photoperiods, at higher latitudes (Behrenfeld et al., 2016; Taylor et al., 1997). By contrast, cooler temperatures at higher latitudes may result in higher chlorophyll-\(a\) to biomass ratios because of lower growth rates, at least when the algae are nutrient-replete (Behrenfeld et al., 2016). Thus, the use of a relative rate (NRCC) as the threshold value for defining a PCI, and as a metric reported in the dataset, facilitates comparisons between lakes of different trophic status or standing stock of chlorophyll-\(a\).

4.2 Chlorophyll-\(a\) rates: temperature and climate warming

The start and end days of the spring and single PCIs show temporal trends towards occurrence earlier in the year (Figure 8a). Earlier springtime algal activity could be linked to global warming. The latter is expected to result in earlier ice break-up and earlier surface water temperature conditions favorable for algal growth (Markelov et al., 2019). The start and end days of the spring PCIs show a positive correlation with increasing temperature (Figure 8b). By contrast, little or even negative correlations are seen for the fall PCIs. Thus, all other conditions unchanged, a warmer climate would see earlier spring blooms, but little temporal shifts for the fall PCIs and, possibly, even a slight delay. For the spring and single PCIs, the duration shows a maximum around 10° C. Therefore, moderate temperatures near or slightly above 10° C should, on average, produce the longest lasting algal growth events. The same trend is not seen for the fall PCIs, possibly because they occur when water temperatures are already above 10° C.

4.3 Surface solar radiation during PCIs: seasonal distributions and distances to lakes

The mean SSR during spring PCIs in the dataset is approximately 100 W m\(^{-2}\) (Figure 9), which is lower than the mean SSR values of single and fall PCIs that are both close to 175 W m\(^{-2}\). This difference in mean SSR between spring and fall PCIs is expected, given the longer daylight hours and more intense sunlight experienced in summer and fall compared to early spring. The similarity in mean SSR between single and fall PCIs may be related to the observation that, at
higher latitudes (>55°N), single PCIs occur more commonly than double PCIs (Figure 4d). Higher latitude lakes tend to bloom only once during the summer months, taking advantage of the period of the year with the highest SSR (Behrenfeld et al., 2016; Lewis, 2011). In support of this, Figures 5b and 5c show that single PCIs tend to occur between late spring and early fall. On the other hand, at lower latitudes (40-45°N), double PCIs are more common than single PCIs, likely due to higher temperatures and longer periods of sufficient daylight experienced during the spring and fall “shoulder seasons” at these latitudes.

Despite the defining importance of sunlight for photosynthesis, in situ SSR time series data are rarely measured systematically as part of lake monitoring programs (Sterner et al., 1997). Although gridded reanalysis datasets that include solar radiation parameters exist, their comparibility with in situ SSR measurements remains in question (Wohland et al., 2020). In gathering open source data, we compiled in situ SSR measurements from locations as close as possible to the lakes with chlorophyll-α data. Nonetheless, much of the SSR values in our dataset were collected at considerable distances from the corresponding lakes (up to ~300 km, Figure 10). For our dataset, only ~10% of the locations where SSR was measured are less than 20 km away from the corresponding lakes, while ~40% are 20-50 km away, ~43% are 50-100 km away, and ~7% are more than 100 km away. Hence, in a significant number of cases, the actual mean SSR during a PCI may differ from the in situ mean SSR reported here, due to differences in cloud cover and levels of atmospheric aerosols (among other factors) (Alpert and Kishcha, 2008). Users are therefore advised to consider this limitation when making use of the SSR values in our dataset. Overall, we recognize a need for SSR data to be more systematically measured and reported as part of lake monitoring programs, in particular for oligotrophic lakes.

5 Conclusions

We present a novel way to delineate annual periods of chlorophyll-α increase (PCIs) in lakes that, presumably, overlap with periods of algal growth. We apply this approach to derive the chlorophyll-α rates of increase (RCIs) during the PCIs of 343 lakes from cold and cold-temperate regions in the northern hemisphere and covering the period 1964-2019. The derived RCIs are assembled in an open-source dataset, together with additional information on the lakes, including water quality, trophic state, and surface solar radiation. Note that the dataset can be paired with other databases, such as HydroLAKES (Messager et al., 2016), HydroATLAS (Linke et al., 2016).
2019), and GLCP (Meyer et al., 2020), to access additional lake and/or watershed attributes. Our dataset is designed to support comparative analyses of the controls on lake chlorophyll-a dynamics and, by extension, also algal dynamics, within and between lakes. We present several examples of such analyses. We hope these will encourage others to use the dataset in their own research and to further expand the dataset’s geographical reach and information content.

**Code and data availability**

All code is available in the project GitHub repository (https://github.com/hfadams/pci) and in Zenodo (https://doi.org/10.5281/zenodo.6972355). The PCI dataset and supplementary data files can be openly accessed at the Federated Research Data Repository at https://doi.org/10.20383/102.0488 (Adams et al., 2021).

**Author contributions**

All authors took part in development of the study. SS, BP, and PVC conceptualized the study, while HA and JY developed the methods and carried out the data collection and data post-processing. HA wrote the original manuscript with contributions from JY, BP, SS, HKP, and PVC. All authors reviewed and edited the final paper.

**Competing interests**

The authors declare that they have no conflict of interest.

**Acknowledgments**

This work was funded by the Lake Futures project within the Global Water Futures (GWF) project supported by the Canada First Research Excellence Fund (CFREF). We would like to thank all the institutions and authors listed in the supplementary information for making their data open source and free to support our work.
References
Adams, H., Ye, J., Slowinski, S., Persaud, B., Kheyrollah Pour, H., and Van Cappellen, P.: Rates and
timing of chlorophyll-a increases and related environmental variables in global temperate and cold-
visualization [version 2; peer review: 2 approved], Wellcome Open Res, 4,
Álvarez, E., Nogueira, E., and López-Urrutia, Á.: In vivo single-cell fluorescence and size scaling of
16/SUPPL_FILE/ZAM999117731S1.PDF, 2017.
Battin, T. J., Kaplan, L. A., Findlay, S., Hopkinson, C. S., Marti, E., Packman, A. I., Newbold, J. D., and
Sabater, F.: Biophysical controls on organic carbon fluxes in fluvial networks, Nat. Geosci., 1, 95–100,
https://doi.org/10.1038/ngeo101, 2008.
Baumert, H. Z. and Petzoldt, T.: The role of temperature, cellular quota and nutrient concentrations for
photosynthesis, growth and light-dark acclimation in phytoplankton, Linnologica, 38, 313–326,
Behrenfeld, M. J., O’Malley, R. T., Boss, E. S., Westberry, T. K., Graff, J. R., Halsey, K. H., Milligan, A.
Cascading trophic interactions and lake productivity, 35, 634–639, 2016.
Integrating Inland Waters into the Terrestrial Carbon Budget, Ecosystems, 10, 171–184,
Kritzberg, E., McKnight, D. M., Freeman, E. C., Senar, O. E., Andersson, A., Ask, J., Berggren, M.,
A.: Global change-driven effects on dissolved organic matter composition: Implications for food webs of
Danielson, J. and Gesch, D.: Global Multi-resolution terrain elevation data 2010 (GMTED2010), U.S.
Deng, J., Paerl, H. W., Qin, B., Zhang, Y., Zhu, G., Jeppesen, E., Cai, Y., and Xu, H.: Climatically-
modulated decline in wind speed may strongly affect eutrophication in shallow lakes, Sci. Total Environ.,
Dorset Environmental Science Centre: Lakeshore Capacity Assessment Handbook: Protecting Water
Quality in Inland Lakes on Ontario’s Precambrian Shield Appendix C, Dorset Environmental Science


Table 1: Summary of variables in the PCI dataset. Associated lake data (e.g., lake depth, surface area, volume, climate zone) are available in the appendix (Table A4).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timing</td>
<td>NA</td>
<td>Three possible PCIs: spring, fall, or single PCI</td>
<td>A single PCI occurs when there is only one maximum in the smoothed yearly chlorophyll-a concentration time series for the year</td>
</tr>
<tr>
<td>Period of chlorophyll-a increase (PCI) start day</td>
<td>Day of year when the PCI begins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Period of chlorophyll-a increase (PCI) end day</td>
<td>Day of year when the PCI ends</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate of chlorophyll-a increase (RCI)</td>
<td>µg L^{-1} day^{-1}</td>
<td>Difference in chlorophyll-a concentration between start and end of the PCI divided by the duration of the PCI</td>
<td>One RCI value is associated with each PCI</td>
</tr>
<tr>
<td>Normalized rate of change in chlorophyll-a (NRCC)</td>
<td>day^{-1}</td>
<td>RCI divided by the initial chlorophyll-a concentration</td>
<td>Accounts for variable standing stock of chlorophyll-a</td>
</tr>
<tr>
<td>Rate of particulate organic carbon (POC) increase</td>
<td>mg L^{-1} day^{-1}</td>
<td>Same calculation as RCI but using start and end POC concentrations</td>
<td>Proxy for the rate of change in total algal biomass</td>
</tr>
<tr>
<td>RCI:rate of POC increase</td>
<td>mg chlorophyll-a mg^{-1} POC</td>
<td>Accounts for variable chlorophyll-a content of algal biomass</td>
<td></td>
</tr>
<tr>
<td>Mean PCI surface water temperature</td>
<td>°C</td>
<td>Mean value during the PCI and the 14-day pre-PCI</td>
<td></td>
</tr>
<tr>
<td>Mean PCI surface solar radiation</td>
<td>W m^{-2}</td>
<td>Mean value during the PCI and the 14-day pre-PCI</td>
<td></td>
</tr>
<tr>
<td>Mean PCI total phosphorus (TP)</td>
<td>mg L^{-1}</td>
<td>(Co-)limiting macronutrients</td>
<td></td>
</tr>
<tr>
<td>Mean PCI soluble reactive phosphorus (SRP)</td>
<td>mg L^{-1}</td>
<td>Mean values during the PCI</td>
<td></td>
</tr>
<tr>
<td>Mean PCI total Kjeldahl nitrogen (TKN)</td>
<td>mg L^{-1}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean PCI Secchi depth</td>
<td>m</td>
<td>Proxy for turbidity</td>
<td></td>
</tr>
<tr>
<td>Mean PCI pH</td>
<td>pH units</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trophic Status Index (TSI)</td>
<td>Range: 0-100</td>
<td>Calculated from chlorophyll-a concentrations across all years the lake was sampled</td>
<td>Basis for assigning trophic status</td>
</tr>
<tr>
<td>Trophic status</td>
<td>NA</td>
<td>Trophic status class assigned based on TSI: Oligotrophic, Mesotrophic, Eutrophic, or Hypereutrophic</td>
<td>TSI thresholds are those of the North American Lake Management Society</td>
</tr>
</tbody>
</table>
Figure 1: Distribution of the 343 lake sampling locations in the PCI dataset. Lake sampling points are clustered by proximity, where marker size and value indicate the number of unique locations represented by each point (light blue markers with white text). Enlarged sections show each lake sampling location (blue markers) and along with the location of the 320 paired SSR stations (orange markers). Base map credit: ESRI, 2011.
Figure 2: Workflow for detecting PCIs and processing data. For each lake sampling point, chlorophyll-a (Chl-a) data are smoothed with a Savitzky-Golay filter and then PCIs are detected based on peaks in the chlorophyll-a concentration. PCIs are flagged as spring, fall, or single PCIs. The data density is shown at key points along the workflow.
Figure 3. Example of spring and fall PCIs in Lake Windermere’s north basin in 1988. The solid grey line is the chlorophyll-a concentration (µg L\(^{-1}\)), and the solid black line is the chlorophyll-a concentration smoothed with a Savitzky-Golay filter. The dashed line is the normalized rate of change in chlorophyll-a (NRCC) (day\(^{-1}\)) where the first derivative is divided by the smoothed chlorophyll-a concentration and is plotted using the right axis. The PCI begins when the NRCC surpasses a threshold of 0.4 day\(^{-1}\) as shown in the first (spring) PCI and ends when the NRCC turns negative, which is when the peak chlorophyll-a concentration is reached. When a peak is detected but the NRCC does not surpass a threshold of 0.4 day\(^{-1}\), the PCI begins when the NRCC surpasses 0 day\(^{-1}\) as shown in the second (fall) PCI. The PCI and pre-PCI (two weeks leading up to the PCI) are shown in dark and light grey shading, respectively.
Figure 4. Distributions of (a) year of occurrence, (b) mean time between samples, (c) lake trophic status index, and (d) lake latitude for each PCI in the dataset. Data are grouped by “double PCI” or “single PCI” year. The data is skewed toward more recent years and higher latitudes. Lakes in the oligotrophic category (TSI < 40) have a higher proportion of single PCIs. These “raincloud plots” show the same data visualized in 3 different ways for each group: frequency distribution, boxplot with quartiles (outliers as represented as points), and a jitter plot of data points as different ways to visualize the data (Allen et al., 2021). Note that the amplitude of the frequency distribution is not proportional between categories.
Figure 5: Frequency Distributions of (a) duration, (b) start day (day of year), and (c) end day (day of year), of the PCIs, grouped by PCI type. Single PCIs have the longest range in length while fall PCIs tend to be the shortest. Single PCIs have the largest range of start and end days while the spring and fall PCIs tend to start and end within a smaller window. These raincloud plots show the same data visualized in 3 different ways for each group: frequency distribution, boxplot with quartiles (outliers represented as points), and a jitter plot of data points.
Figure 6. Distributions of selected water quality variables during PCIs: (a) log rate of chlorophyll-a increase, (b) mean surface water temperature, (c) log mean total phosphorus (TP), and (d) mean Secchi depth. The mean rate of chlorophyll-a increase is lowest in the single PCI category and highest in the fall PCIs. For the single PCIs, temperature is evenly distributed across the annual range as they occur throughout the ice-free season. Total phosphorus concentrations are lowest during the spring PCIs, which likely reflects a greater control of P limitation on algal growth during spring compared to summer and fall. Each PCI category has a similar range in Secchi depth, between 0 and 5 m. Raincloud plots show the frequency distribution, boxplot with quartiles (outliers as represented as points), and a jitter plot of data points for each group.
Figure 7. Rate of Chlorophyll-a increase (RCI) trends in the dataset, grouped by (a) trophic status, (b) latitude, and (c) climate zone. Lakes of a higher trophic status have a higher mean RCI while lakes at higher latitudes have lower RCI (with considerable overlap between all categories). Grouping by climate zone shows minimal effect on RCI. The number of lakes represented by each violin is shown in grey text on the panels. Climate zones are as follows: 7 = cold and mesic; 8 = cool, temperate, and dry; 10 = warm, temperate, and mesic. White circles indicate the mean value for each violin.
Figure 8. (a) Start and end days for the spring, fall, and single PCIs for all the lakes in the dataset; spring and single PCI categories trend toward earlier start and end days, while fall PCI start days are occurring earlier in the year. (b) Start and end days of the PCIs as a function of temperature (top two rows in panel B, linear regression trendline in black) suggest a positive relationship between PCI timing and surface water temperature in the spring and a negative relationship in the fall. Longer PCIs occur at moderate surface water temperatures which are observed less often during the fall PCIs (trendline fitting data in the bottom row is locally weighted scatterplot smoothing).
Figure 9: Mean PCI surface solar radiation (SSR) grouped by PCI type (single, spring, or fall). White circles show the mean value for each violin. The mean SSR during spring PCIs is lower than that of single and fall PCIs, which have similar distributions.
Figure 10: Frequency distribution of distances between the lake sampling points and the nearest surface solar radiation (SSR) sampling stations, in decimal degrees. Most lake-SSR distances are within 200 kilometres of each other. Cloud cover, atmospheric aerosols, and their interactions are a major control on incident SSR at a given surface location, therefore, the SSR values may become less representative of the paired lake with increasing distances. The middle line in the boxplot shows the median value.
## Appendix A: Data sources and additional lake attributes

Table A1: Summary of sources and licensing for the chlorophyll-a data. Direct links to the datasets are provided where possible, and lake names can be searched within the database. Note that not all lakes in these databases met the requirements to be retained in the PCI dataset.

<table>
<thead>
<tr>
<th>Database</th>
<th>Region</th>
<th>Lake(s)</th>
<th>Data license</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Canada</td>
<td>Ontario, Manitoba</td>
<td>Laurentian great lakes, Hamilton Harbour, Riding Mountain lakes</td>
<td>Open Government Licence</td>
</tr>
<tr>
<td>Lake Winnipeg DataStream</td>
<td>Manitoba</td>
<td>Lake Winnipeg</td>
<td>Open Government Licence</td>
</tr>
<tr>
<td>CanWIN Data HUB</td>
<td>Manitoba</td>
<td>Lake Winnipeg</td>
<td>Open Data</td>
</tr>
<tr>
<td>IISD-ELA private database</td>
<td>Ontario</td>
<td>Experimental lakes 114, 224, 239, and 442</td>
<td>Terms and conditions</td>
</tr>
<tr>
<td>Alberta Environment and Parks data repository</td>
<td>Alberta</td>
<td>Many lakes sampled by Alberta Environment and Parks</td>
<td>Open Government Licence</td>
</tr>
<tr>
<td>LUBW data and map service</td>
<td>Germany</td>
<td>Constance Untersee</td>
<td>User agreement</td>
</tr>
<tr>
<td>Water Information System Sweden (VISS)</td>
<td>Sweden</td>
<td>Hundreds of lakes monitored across Sweden</td>
<td>CC0 license - free use</td>
</tr>
<tr>
<td>UK Environment Agency</td>
<td>UK</td>
<td>Many lakes monitored across the UK</td>
<td>Terms of use</td>
</tr>
<tr>
<td>UK Centre for Ecology and Hydrology</td>
<td>UK</td>
<td>Bassenthwaite, Belhalm tarn, Derwent water, Esthwaite water, Grasmere,</td>
<td>Open Government Licence v3,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loch leven, Lake Windermere (north basin), Lake Windermere (south basin)</td>
<td>Terms of use</td>
</tr>
<tr>
<td>Environmental Data Initiative portal</td>
<td>Global</td>
<td>Central long lake, East long lake, Giles lake, Lacawac, May lake, Paul</td>
<td>Creative Commons license CC-BY 4.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lake, Peter lake, Tuesday lake, Wavynwood lake, West long lake</td>
<td></td>
</tr>
<tr>
<td>Knowledge Network for Biocomplexity (KNB)</td>
<td>Colorado</td>
<td>Oneida lake</td>
<td>Open Data Commons Attribution</td>
</tr>
<tr>
<td>University of Wisconsin NLTER</td>
<td>Wisconsin</td>
<td>Allequash lake, Big Muskellunge lake, Crystal lake, Crystal bog, Sparkling</td>
<td>Data use agreement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lake, Trout lake, Trout bog, Fish lake, Lake Mendota, Lake Monona, Lake</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wingra</td>
<td></td>
</tr>
<tr>
<td>USGS and USEPA water quality</td>
<td>USA</td>
<td>USGS-491528094470601, USGS-49212094421501</td>
<td>User guide</td>
</tr>
</tbody>
</table>
Table A2: Summary of sources and licensing for the surface solar radiation data. Direct links to the databases are provided where possible, but the Environment and Climate Change Canada (ECCC) and IISD-ELA data were acquired through communication with the curators.

<table>
<thead>
<tr>
<th>Database</th>
<th>Region</th>
<th>Data license</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETH Zurich GEBA</td>
<td>Global</td>
<td>Data availability</td>
<td>Data provided by Alberta Agriculture and Forestry, and Alberta Climate Information Service (ACIS) (August 2020)</td>
</tr>
<tr>
<td>Agriculture AB Station Data</td>
<td>Alberta</td>
<td>Terms of use</td>
<td>DOI: <a href="https://doi.org/10.1594/PANGAEA.880000">https://doi.org/10.1594/PANGAEA.880000</a></td>
</tr>
<tr>
<td>Baseline Solar Radiation Network</td>
<td>Global</td>
<td>Creative Commons license CC-BY 4.0</td>
<td>Source: direct communication with ECCC</td>
</tr>
<tr>
<td>Environment and Climate Change Canada (ECCC)</td>
<td>Canada</td>
<td>Open Government License</td>
<td>Source: direct communication with ECCC</td>
</tr>
<tr>
<td>IISD-ELA private database</td>
<td>Ontario</td>
<td>Terms and conditions</td>
<td>Source: direct communication with IISD-ELA</td>
</tr>
</tbody>
</table>

Table A3: Summary of miscellaneous databases used to acquire lake attributes. Follow embedded links to access the database and metadata.

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Multi-resolution Terrain Elevation Data (GMTED2010)</td>
<td>Global digital elevation model used to extract lake and SSR station elevation in QGIS (Danielson and Gesch, 2010)</td>
</tr>
<tr>
<td>HydroLAKES V1.0</td>
<td>Global lake shapefile used to assign lake area, mean depth, and volume (ver. 1.0; Messager et al., 2016)</td>
</tr>
<tr>
<td>HydroATLAS V1.0</td>
<td>Global shapefile for regional attributes, used to assign climate zone to each lake (ver. 1.0; Linke et al., 2019)</td>
</tr>
</tbody>
</table>
### Table A4: Lake attributes published alongside the PCI dataset (https://doi.org/10.20383/102.0488)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lake</td>
<td>NA</td>
<td>Lake name, reformatted from original file</td>
</tr>
<tr>
<td>lake_lat</td>
<td>decimal degrees</td>
<td>Lake latitude, collected from original data files and HydroLAKES data (Messager et al., 2016)</td>
</tr>
<tr>
<td>lake_long</td>
<td>decimal degrees</td>
<td>Lake longitude, collected from original data files and HydroLAKES data (Messager et al., 2016)</td>
</tr>
<tr>
<td>tsi</td>
<td>range from 0-100</td>
<td>Calculated from mean chlorophyll-a concentration across all years the lake was sampled, based on guidelines from the North American Lake Management Society</td>
</tr>
<tr>
<td>trophic_status</td>
<td>oligotrophic, mesotrophic, eutrophic, hypertrophic</td>
<td>Assigned using lake trophic status index</td>
</tr>
<tr>
<td>climate_zone</td>
<td>integer</td>
<td>Climate zone of each lake, assigned using the HydroATLAS database (Linke et al., 2019)</td>
</tr>
<tr>
<td>lake_elev</td>
<td>m above sea level</td>
<td>Elevation of the lake, extracted from the Global Multi-resolution Terrain Elevation Data (GMTED2010) model (Danielson and Gesch, 2010)</td>
</tr>
<tr>
<td>lake_area</td>
<td>km²</td>
<td>Total lake surface area, extracted from the HydroLAKES database (Messager et al., 2016)</td>
</tr>
<tr>
<td>lake_volume</td>
<td>km³</td>
<td>Total lake volume, extracted from the HydroLAKES database (Messager et al., 2016)</td>
</tr>
<tr>
<td>mean_lake_depth</td>
<td>m</td>
<td>Mean lake depth, extracted from the HydroLAKES database (Messager et al., 2016)</td>
</tr>
<tr>
<td>start_sampling</td>
<td>year</td>
<td>Year when lake sampling started</td>
</tr>
<tr>
<td>end_sampling</td>
<td>year</td>
<td>Year when lake sampling ended</td>
</tr>
<tr>
<td>days_sampled</td>
<td>days</td>
<td>Total number of days where lake data were recorded in the original dataset</td>
</tr>
<tr>
<td>years_sampled</td>
<td>years</td>
<td>Total number of years where lake data were recorded in the original dataset</td>
</tr>
<tr>
<td>mean_time_between_samples</td>
<td>days</td>
<td>Average number of days between sample collection (sampling resolution)</td>
</tr>
<tr>
<td>lake_data_source</td>
<td>NA</td>
<td>Name of database where the original lake data was sourced</td>
</tr>
<tr>
<td>country</td>
<td>NA</td>
<td>Name of the country where the lake is located</td>
</tr>
<tr>
<td>variables</td>
<td>NA</td>
<td>List of the variables found in the pci_dataset/PCI dataset for each lake</td>
</tr>
<tr>
<td>ssr_station</td>
<td>NA</td>
<td>Station name as assigned in original database</td>
</tr>
<tr>
<td>ssr_id</td>
<td>NA</td>
<td>ID number in original database (where available)</td>
</tr>
<tr>
<td>ssr_source</td>
<td>NA</td>
<td>Name of database where the original SSR data was sourced</td>
</tr>
<tr>
<td>ssr_lat</td>
<td>decimal degrees</td>
<td>SSR station latitude</td>
</tr>
<tr>
<td>ssr_long</td>
<td>decimal degrees</td>
<td>SSR station longitude</td>
</tr>
<tr>
<td>geo_dist_km</td>
<td>km</td>
<td>Geodesic distance between the paired lake and SSR station</td>
</tr>
<tr>
<td>ssr_elev</td>
<td>m above sea level</td>
<td>Elevation of the SSR station, extracted from the Global Multi-resolution Terrain Elevation Data (GMTED2010) model (Danielson and Gesch, 2010)</td>
</tr>
<tr>
<td>ssr_lake_elev_diff</td>
<td>m</td>
<td>Difference in elevation between the paired lake and SSR station (positive when the SSR station is at a higher elevation)</td>
</tr>
<tr>
<td>ssr_start</td>
<td>year</td>
<td>Year when SSR sampling started</td>
</tr>
<tr>
<td>ssr_end</td>
<td>year</td>
<td>Year when SSR sampling ended</td>
</tr>
<tr>
<td>ssr_years_sampled</td>
<td>years</td>
<td>Total number of years where SSR data were recorded in the original dataset</td>
</tr>
<tr>
<td>ssr original resolution</td>
<td>month or day</td>
<td>Resolution of the original SSR data before it was resampled to achieve a daily resolution</td>
</tr>
</tbody>
</table>